Testing the mechanism of missing data

Denys Pommeret

Testing the mechanism of missing data

Denys Pommeret (pommeret@amu-univ.fr)
Institute of Mathematics, Aix Marseille University

Abstract

We consider the problem of missing data when the mechanism of missingness is not at random and when the partially observed variable has known or observed moments. A nonparametric estimator of the probability of missingness is proposed. A data driven statistic is constructed to test the missingness mechanism. Illustrations through univariate logistic regressions are presented: the method permits to estimate regression coefficients when the covariate is completely missing for one response category. A test of significance is proposed for the coefficients. The performance of the method is investigated in a simulation study. An illustration is considered using a real data set.

1 INTRODUCTION

Missing data are frequently encountered in data analysis and the missingness may sometimes depend on the unobserved value. For example if variables are connected to personal information as the income, the quality of life, or the political opinion. In this case the mechanism is classified as Missing Not At Random, according to Rubin (1987). We study this mechanism when the moments of the missing variable are known or estimated. This knowledge is realistic when the distribution of the variable is known in the entire population or observed independently. We will denote by Y a non null univariate random variable and by W a missing indicator taking value 0 if Y is missing and 1 otherwise. We consider the case where the probability that $W = 0$ depends of Y and we write

$$W|Y \quad = \quad \begin{cases} 1 & \text{with probability} \quad p(Y) = \mathbb{P}(W = 1|Y), \\ 0 & \text{with probability} \quad 1 - p(Y) = \mathbb{P}(W = 0|Y). \end{cases}$$

It is assumed that $X = WY$ is observed. We propose an estimator of the probability of $W = 1$. The main idea is that the probability is a bounded

function of Y that can be expressed in an orthonormal basis $\mathcal{B}$. Moreover, the basic property $\mathbb{E}(X^n) = \mathbb{E}(WY^n) = \mathbb{E}(p(Y)Y^n)$ orients the choice of $\mathcal{B}$ to a basis of orthonormal polynomials. This expansion of p is then used to construct a test to determine the randomness of the missing data mechanism; that is, to test if W and Y are independent. The test statistic is inspired by the work of Neyman (1937) (see more recently Rayner and Best, 1989, 2001) and can be constructed in the same spirit of Ignaccolo (2004).

As an application, we consider the univariate logistic regression model with two coefficients a and b and with covariate Y observed only when the response $W = 1$ and such that

$$\text{logit}\big(\mathbb{P}(W = 1|Y))\big) \quad = \quad aY + b.$$

This situation may correspond to the case where we obtain information on the consumers of a product, as the age of an insurant, or as the income of a customer of a bank, being customer corresponding to $W = 1$ here. The previous approach can be adapted with two objectives: to estimate the regression coefficients by using the estimator of p, and to test their significant by testing the mechanism of missingness. We develop this model in a simulation study. We also illustrate this situation through a survey collected for the French National Institute of Statistics and Economic Studies (INSEE).

The rest of the paper is organized as follows. In Section 2 we develop the construction of an estimator of missingness probabilities. In Section 3 we proceed with the construction of the test. Section 4 is devoted to the simulation and Section 5 presents a real case study.

2 ESTIMATING THE MISSINGNESS PROBABILITIES

2.1 The case where the distribution of Y is known

Let $X_1, \cdots, X_n$ be i.i.d. univariate random variables such that

$$X_i \quad = \quad Y_i W_i,$$

where Y_i are i.i.d. non null random variables and where $W_i|Y_i = 1$ with probability $p(Y_i)$ and 0 with probability $1 - p(Y_i)$

Remark 1. *If Y is a discrete random variable on $\mathbb{N}$ we can adapt our model, obtaining $Y > 0$ by translation.*

The data are missing not at random except when the probability p is constant. We assume first that the random variables Y_i have known probability measure μ on a support S. We denote by $\mathcal{B} = \{Q_n; n = 0, 1, \cdots\}$ an associated basis of dense orthonormal polynomials with respect to μ; that is, each Q_j is of degree j and

$$\int_S Q_i(x)Q_j(x)\mu(dx) \quad = \quad \delta_{ij},$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. We put $Q_0 = 1$.

Proposition 1. *For all $y \in S$, we have:*

$$p(y) \quad = \quad \mathbb{E}(W) + \sum_{k>0} \big\{\mathbb{E}(Q_k(X)) + Q_k(0)(\mathbb{E}(W) - 1)\big\}Q_k(y).$$

Proof Since p is bounded it satisfies the following expansion

$$\begin{aligned} p(y) &= \sum_{k\in\mathbb{N}} \int_{\mathcal{S}} p(t)Q_k(t)\mu(dt)Q_k(y) \\ &= \sum_{k\in\mathbb{N}} \mathbb{E}\big(Q_k(Y)p(Y)\big)Q_k(y). \end{aligned}$$

Using the equalities $\mathbb{E}(X^k) = \mathbb{E}(Y^k W) = \mathbb{E}(Y^k p(Y))$, for all $k \in \mathbb{N}^*$, we get

$$\begin{aligned} \mathbb{E}\big(Q_k(X)\big) &= \mathbb{E}\big\{Q_k(X) - Q_k(0)\big\} + Q_k(0) \\ &= \mathbb{E}\big\{(Q_k(Y) - Q_k(0))p(Y)\big\} + Q_k(0), \end{aligned}$$

and observing that $\mathbb{E}(p(Y)) = \mathbb{E}(W)$ we deduce the result. ■

Then, for a given integer $K > 0$, a Kth order estimator of the probability $p(y)$ is given by

$$\widehat{p}_K(y) \quad = \quad \sum_{k\le K} \big\{E_k + Q_k(0)(C - 1)\big\}Q_k(y),$$

where

$$E_k = \frac{1}{n}\sum_{i=1}^{n} Q_k(X_i), \quad \text{and} \quad C = \frac{1}{n}\sum_{i=1}^{n} W_i.$$

We consider the MISE (Mean Integrated Square Error) criterion to evaluate the behavior of this estimator. Write

$$\begin{aligned} p_K(y) &= \sum_{k\leq K} \big\{\mathbb{E}(Q_k(X)) + Q_k(0)(\mathbb{E}(W)-1)\big\}Q_k(y), \\ \|p\|_\mu^2 &= \int_{\mathcal{S}} p(y)^2 \mu(dy). \end{aligned}$$

Proposition 2. *We have*

$$\mathbb{E}\big(\|p-\widehat{p}_K\|_\mu^2\big) \leq \|p-p_K\|_\mu^2 + \frac{K+1}{n}.$$

Proof From the orthogonality of the polynomials we have

$$\begin{aligned} \mathbb{E}\big(\|p-\widehat{p}_K\|_\mu^2\big) &= \|p-p_K\|_\mu^2 + \mathbb{E}\big(\|p_K-\widehat{p}_K\|_\mu^2\big) \\ &= \|p-p_K\|_\mu^2 + \frac{1}{n}\sum_{i\leq K} \mathbb{V}(Q_i(X)+Q_i(0)(W-1)) \\ &= \|p-p_K\|_\mu^2 + \frac{1}{n}\sum_{i\leq K} \mathbb{V}(Q_i(Y)W), \end{aligned}$$

and we use the inequality $\mathbb{V}(Q_i(Y)W) \leq 1$ to conclude. ■

2.2 The case where the distribution of Y is estimated

Here we assume that only $\mathcal{S}$, the support of μ, is known and we denote by f its density with respect to a given measure ν. We consider a basis $\{P_i, i=0,1,2,\cdots\}$ of ν orthonormal polynomials. We obtain the following adaptation of Proposition 1.

Proposition 3. *Assume that $f>0$ and that $\int_{\mathcal{S}} f(x)^2\nu(dx) < \infty$. Then we have for all $y \in \mathcal{S}$*

$$p(y) = \frac{1}{f(y)} \sum_{k\geq 0} \big(\mathbb{E}(P_k(X)) + P_k(0)(\mathbb{E}(W)-1)\big)P_k(y).$$

Proof The proof is similar to the Proposition 1. ■

If Y is observed through another independent sample of size N, say $U_1, \cdots, U_N$, we can estimate f via a kernel $\mathcal{K}(\cdot)$ and use Proposition 3 to construct an estimator of p. Define

$$\widehat{p}_{K,\widehat{f}}(y) \quad = \quad \frac{1}{\widehat{f}(y)} \sum_{k \leq K} \big(\tilde{E}_k + P_k(0)(C-1)\big) P_k(y),$$

where

$$\tilde{E}_k = \frac{1}{n}\sum_{i=1}^{n} P_k(X_i),\ C = \frac{1}{n}\sum_{i=1}^{n} W_i,$$
$$\text{and } \widehat{f}(y) = \min\big(e_N, \frac{1}{Nh_N}\sum_{i=1}^{N} \mathcal{K}\left(\frac{U_i - y}{h_N}\right)\big),$$

with an appropriate bandwidth h_N and a trimming $e_N \to 0$ as $N \to \infty$. Under basic assumptions we have the following extension of Proposition 2.

Proposition 4. *Assume that $f > 0$ and f is $\mathcal{C}^d$, for some positive integer d. If $h_N \simeq N^{-c_1}$, $e_N \simeq N^{-c_2}$ for some positive constants c_1 and c_2 such that $\frac{2c_2}{d} < c_1 < 1 - 4c_2$, and if $N = o(n^{1/2c_2})$, then under H_0 we have*

$$\mathbb{E}\big(\|p - \widehat{p}_{K,\widehat{f}}\|_\mu^2\big) \quad = \quad \|p - p_{K,f}\|_\mu^2 + o(1).$$

Proof Let us evaluate

$$\begin{aligned}\mathbb{E}\big(\|\widehat{p}_{K,\widehat{f}} - p\|_\mu^2\big) &= \mathbb{E}\big(\|\widehat{p}_{K,\widehat{f}} - \widehat{p}_{K,f}\|_\mu^2\big) + \mathbb{E}\big(\|\widehat{p}_{K,f} - p\|_\mu^2\big) \\ &\quad -2\mathbb{E}\big(\langle \widehat{p}_{K,\widehat{f}} - p\,,\, \widehat{p}_{K,\widehat{f}} - \widehat{p}_{K,f}\rangle\big) \\ &= (A) + (B) - (C).\end{aligned}$$

Write $\gamma_k = \tilde{E}_k + P_k(0)(C-1)$ and $g_k = \mathbb{E}(\gamma_k)$. It is clear that $\mathbb{E}(\gamma_k^2) < \infty$ and that $\mathbb{E}(\gamma_k - g_k)^2 = O(1/n)$. By a first order expansion in (C) we get

$$\begin{aligned}|(C)| \quad \leq \quad & 2\sup|\widehat{f} - f| \frac{1}{e_N^2} \sum_{i,j=1}^{k} \mathbb{E}(|\gamma_i\gamma_j|) \int |P_i(y)P_j(y)|\nu(dy) \\ & +2\sup|\widehat{f} - f| \frac{1}{e_N^2} \sum_{i=1}^{k} \mathbb{E}(|\gamma_i|) \int |P_i(y)|p(y)\mu(dy).\end{aligned}$$

Combining the fact that $\int |P_i(y)|p(y)\mu(dy) \leq \int f^2(y)\nu(dy) < \infty$ and that $\int |P_i(y)P_j(y)|\nu(dy) \leq 1$ with the following fundamental property (see for

instance Härdle, 1992): $\sup|\hat{f}-f| = O\left(h_N^d + \sqrt{\log N/Nh_N}\right)$, we obtain $|(C)| = o(1)$. We now examine (B):

$$\begin{aligned}(B) &= \mathbb{E}\big(\|p_{K,f}-p\|_\mu^2\big) + \mathbb{E}\big(\|\widehat{p}_{K,f}-p_{K,f}\|_\mu^2\big) - 2\mathbb{E}\big(\langle p_{K,\widehat{f}}-p\,,\,\widehat{p}_{K,f}-p_{K,f}\rangle\big)\\ &= \mathbb{E}\big(\|p_{K,f}-p\|_\mu^2\big) + B2 + B3.\end{aligned}$$

We continue in this fashion obtaining

$$\begin{aligned}|B3| &\leq 2\sum_{i,j=1}^{K}\mathbb{E}|\gamma_i(\gamma_j-g_j)|\frac{1}{e_N} + 2\sum_{j=1}^{K}\mathbb{E}|\gamma_j-g_j|\frac{1}{e_N}\\ &= o(1).\\ |B2| &\leq \sum_{i,j=1}^{K}\mathbb{E}|(\gamma_i-gi)(\gamma_j-g_j)|\frac{1}{e_N}\\ &= o(1).\\ |(A)| &\leq \sup|\widehat{f}-f|\frac{1}{e_N^2}\sum_{j=1}^{K}\mathbb{E}|\gamma_j|\int f^2(y)\nu(dy)\\ &= o(1).\end{aligned}$$

■

3 TESTING THE MISSING MECHANISM

3.1 The case where the distribution of Y is known

We assume that the distribution of Y is known, with finite moments of all orders, and that $\mathbb{E}(W) > 0$. We consider the following hypotheses

$$\begin{aligned}H_0 &: Y \text{ and } W \text{ are independent}\\ H_1 &: Y \text{ and } W \text{ are dependent.}\end{aligned}$$

The null hypothesis corresponds to a Missing At Random mechanism. With the notation $w = \mathbb{E}(W)$, testing H_0 is equivalent to test $p(y) = w$ for all $y \in \mathcal{S}$. From Proposition 1 it is equivalent to test the equalities $\mathbb{E}(Q_k(X)) = Q_k(0)(1-w)$, for $k = 1, 2, \cdots$. Thus, under the null hypothesis our method consists in comparing the estimators associated to $\mathbb{E}(Q_k(X))$ with the estimators associated to $Q_k(0)(1-w)$. Write

$$\alpha_k = 1/\sqrt{n}\sum_{i=1}^{n}Q_k(X_i), \quad \beta_k = 1/\sqrt{n}\sum_{i=1}^{n}Q_k(0)(1-W_i),$$
$$\text{and} \quad U_k = (\alpha_1-\beta_1, \cdots, \alpha_k-\beta_k).$$

By the Central Limit Theorem we have the following convergence in law under H_0:

$$U_k \quad \longrightarrow_{\mathcal{L}} N(0, \Sigma_k),$$

where Σ_k is the $k \times k$ covariance matrix of U_k. Is is easily seen that its (i,j)th element is

$$\Sigma_k(i,j) \quad = \quad \mathbb{V}\big(Q_i(X)Q_j(X)\big),$$

and combining the decomposition $Q_i(X) = (Q_i(Y) - Q_i(0))W + Q_i(0)$ with the orthogonality of the polynomials we finally obtain under H_0 that

$$\Sigma_k \quad = \quad wI,$$

where I denotes the identity matrix. Then under H_0 the statistic $T_k = \|U_k\|^2$ converges to $w\{V_1 + \cdots + V_k\}$, where $V_1, \cdots, V_k$ are independent Chi-square random variables of degree 1. Such a decomposition has previously been used in Ignaccolo (2004). To select the number k of components in the test statistic we follow the work of Kallenberg and Ledwina (1995) (see also Ledwina, 1994) and we consider an increasing sequence of number of components $k(n)$ such that $\lim_{n\to\infty} k(n) = \infty$. The selection rule is based on the following Schwarz's criteria (1978)

$$S_n = \min\Big\{ \underset{1\le k\le k(n)}{\operatorname{argmax}} \left(T_k - k\log(n)\right)\Big\},$$

and the associated data driven test statistic is T_{S_n}. We will need the following assumption:

(A1) $k(n) = o(\sqrt{\log(n)})$.

Theorem 1. *Let assumption* $(A1)$ *holds. Then, under* H_0, T_{S_n}/C *converges in distribution to a Chi-squared random variable with degree 1.*

Proof Under H_0, T_1 converges to a scaled Chi-squared random variable with one degree of freedom and with scale parameter w. As $w > 0$, for n large enough $C > 0$ almost surely and T_1/C converges to the expected distribution. The procedure is then to show that $\mathbb{P}(S_n \ge 2)$ tends to zero. Since $(U_k = k)$ implies $(T_k - k\log(n) \ge T(1) - \log(n))$ we have $\mathbb{P}(U_k = k) \le \mathbb{P}(T_k > (k-1)\log(n))$ and we obtain

$$\mathbb{P}(S_n \ge 2) = \sum_{k=2}^{k(n)} \mathbb{P}(S_n = k) \le \sum_{k=2}^{k(n)} \mathbb{P}\Big(T_k^{1/2} \ge \sqrt{(k-1)\log(n)}\Big).$$

From Markov's inequality we get

$$\mathbb{P}\Big(T_k^{1/2} \geq \sqrt{(k-1)\log(n)}\Big) \quad \leq \quad \frac{\Big(\mathbb{E}\big(\|U_k\|^2\big)\Big)^{1/2}}{\big((k-1)\log(n)\big)^{1/2}}.$$

Using the independence of the pairs $(X_s, Y_s)_{1\leq s\leq n}$, we have

$$\mathbb{E}\big(\|U_k\|^2\big) \quad = \quad k\Big(\frac{1}{k}\sum_{i=1}^{k}\mathbb{V}\big(Z_i\big)\Big),$$

where $Z_i = Q_i(X) - Q_i(0)(1-W)$. An easy computation shows that under H_0, $\mathbb{V}(Z_i) = 1 + Q_i(0)^2\mathbb{E}(1-W)^2$ and then

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{V}\big(Z_i\big) \leq \frac{1}{k}\sum_{i=1}^{k}(1+Q_i(0)^2) < M,$$

where M is a constant determined by the choice of the basis $\mathcal{B}$. Finally, we have

$$\mathbb{P}(S_n \geq 2) \quad \leq \quad \frac{\sqrt{2M}k(n)}{\sqrt{\log(n)}},$$

which gives the result. ■

3.2 The case where the moments of Y are known or estimated

We first assume that the moments of Y are known and that $w = \mathbb{E}(W) > 0$. We use the same notation as in Section 2.2. From Proposition 3,

$$p(y) = w, \forall y \in S \quad \Leftrightarrow \quad wf(y) = \sum_{k\geq} c_k P_k(y), \forall y \in S,$$

with $c_k = \mathbb{E}(P_k(X)) + P_k(0)(\mathbb{E}(W) - 1)$. This can be rewritten as

$$p(y) = w, \forall y \in S \quad \Leftrightarrow \quad w\sum_{k\geq}\mathbb{E}(P_k(Y))P_k(y) = \sum_{k\geq} c_k P_k(y), \forall y \in S,$$

and then H_0 is equivalent to the following equalities

$$w\big(\mathbb{E}(P_k(Y)) - P_k(0)\big) = \mathbb{E}(P_k(X)) - P_k(0), \qquad \forall k > 0.$$

Writing

$$\tilde{\alpha}_k = 1/\sqrt{n}\sum_{i=1}^{n} W_i \mathbb{E}(P_k(Y)) \quad \text{and} \quad \tilde{\beta}_k = 1/\sqrt{n}\sum_{i=1}^{n} \big(P_k(X_i) + P_k(0)(W_i - 1),$$

we can proceed analogously to the previous case, replacing α and β by $\tilde{\alpha}$ and $\tilde{\beta}$, respectively. We consider $\tilde{U}_k = (\tilde{\alpha}_1 - \tilde{\beta}_1, \cdots, \tilde{\alpha}_k - \tilde{\beta}_k)$, $\tilde{T}_k = \|\tilde{U}_k\|^2$, and $\tilde{S}_n = \min\big\{\operatorname{argmax}_{1\le k\le k(n)}(\tilde{T}_k - k\log(n))\big\}$. The test statistic is now $\tilde{T}_{\tilde{S}_n}$. We assume that there exists a convergent estimator of $\mathbb{V}(U_1) > 0$, denoted by $V > 0$. We can now rephrase Theorem 1 as follows.

Theorem 2. *Let Assumptions (A1) holds. Then, under H_0, $\tilde{T}_{\tilde{S}_n}/V$ converges in distribution to a Chi-squared random variable with degree 1.*

Eventually, if the moments of μ are unknown, but estimated through another independent sample of size N, say $U_1, \cdots, U_N$, where U_i are i.i.d. random variables with distribution μ, we can generalize the previous result. For $k = 1, 2, \cdots$, we denote by $e_{k,N}$ a convergent estimator of $\mathbb{E}(P_k(Y))$ based on the sample of size N and we write

$$\alpha'_k = 1/\sqrt{n}\sum_{i=1}^{n} W_i e_{k,N}, \quad \text{with} \quad e_{k,N} = 1/N\sum_{i=1}^{N} \mathbb{E}(P_k(U_i)).$$

Then replacing $\tilde{\alpha}$ by α' in the statistic $\tilde{T}_{\tilde{S}_n}$ we can state the analogue of Theorem 2. We consider $U'_k = (\alpha'_1 - \tilde{\beta}_1, \cdots, \alpha'_k - \tilde{\beta}_k)$, $T'_k = \|U'_k\|^2$, and $S'_n = \min\big\{\operatorname{argmax}_{1\le k\le k(n)}(T'_k - k\log(n))\big\}$. The test statistic is now $T'_{S'_n}$.

Theorem 3. *Let Assumptions (A1) holds and assume that $N = O(n)$. Then, under H_0, $T'_{S'_n}/V$ converges in distribution to a Chi-squared random variable with degree 1.*

4 SIMULATIONS THROUGH A LOGISTIC REGRESSION

4.1 Estimation of the missing probability

We simulated n i.i.d. random variables $Y_1, \cdots, Y_n$ with standard normal distribution $\mathcal{N}(0, 1)$. We constructed $(X_1, \cdots, X_n)$ such that $X_i = Y_i W_i$, where

$$p(y) = \mathbb{P}(W_i = 1 | Y_i = y) = \exp\{ay + b\}/(1 + \exp\{ay + b\}). \tag{1}$$

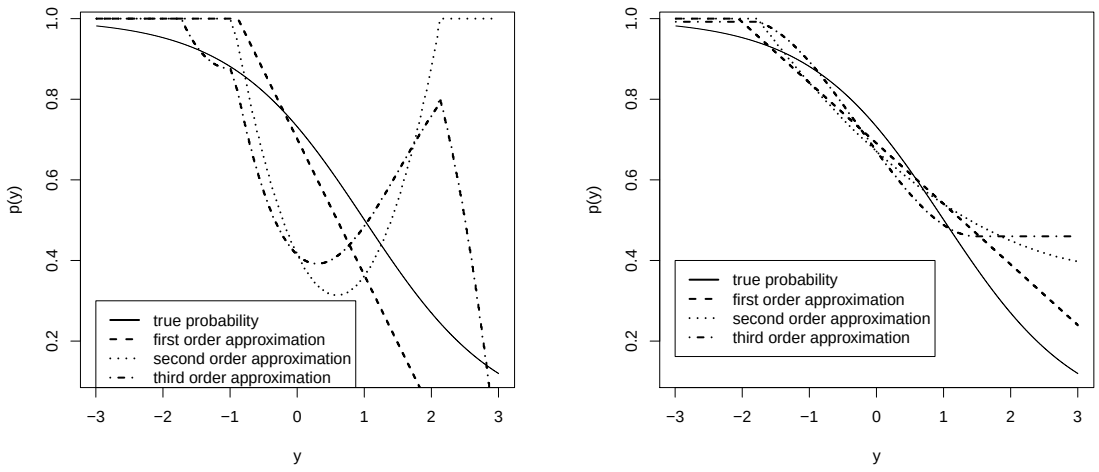

Figure 1: Probability p and its estimates of first, second and third order for the logistic model with coefficients $a = -1, b = 1$ and for two samples of size $n = 50$ (left) and $n = 100$ (right) respectively.

For simulations we chose $a = -1, 0, 1$, $b = 1$, and $n = 50, 100$. Since μ is the normal distribution, the associated orthonormal polynomials are the Hermite ones (see for instance Abramowitz and Stegun, 1972). The three first terms are $P_0 = 1$, $P_1(x) = x$ and $P_2(x) = x^2 - 1$.

Figures 1-3 show the probability p and its estimates of first, second and third order. For small sample size ($n = 50$) the estimates were unstable. This instability is illustrated on the left of both Figures 1 and 2, although Figure 3 shows a case with good estimates. In the case of small sample size it may be preferable to retain the first approximation often more stable.

For a larger sample size ($n = 100$) we got better estimates, close to the probability p, except for the values of y close to -3 or 3 which are rarely observed since Y is $\mathcal{N}(0, 1)$ distributed.

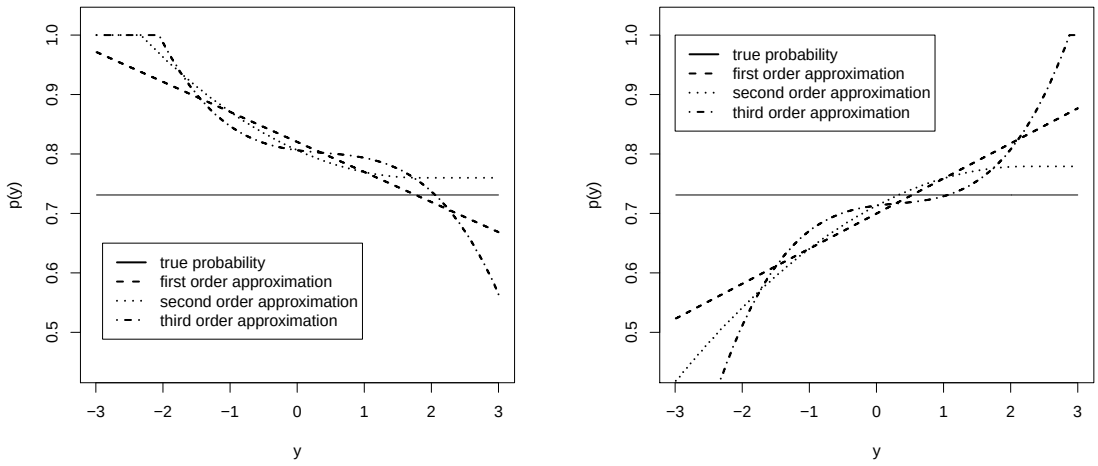

Figure 2: Probability p and its estimates of first, second and third order for the logistic model with coefficients $a = 0, b = 1$ and for two samples of size $n = 50$ (left) and $n = 100$ (right) respectively.

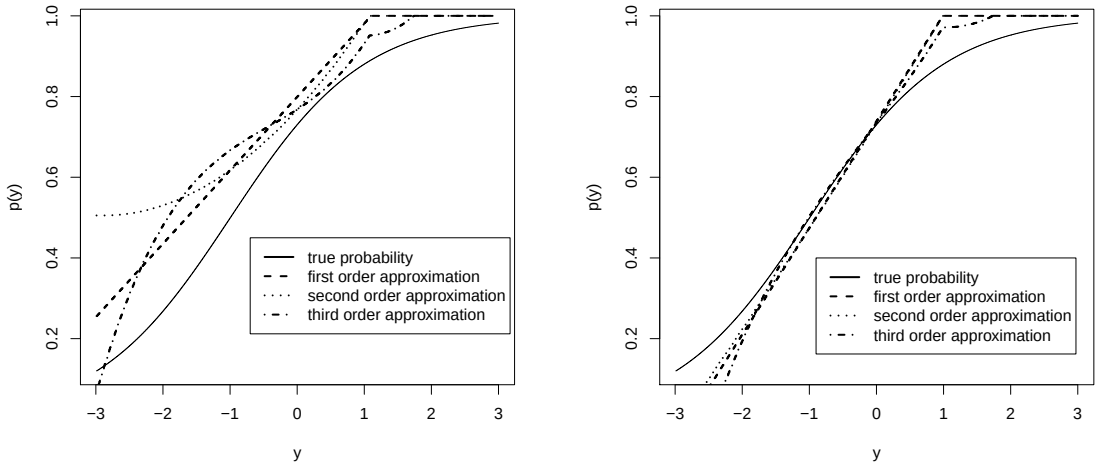

Figure 3: Probability p and its estimates of first, second and third order for the logistic model with coefficients $a = 1, b = 1$ and for two samples of size $n = 50$ (left) and $n = 100$ (right) respectively.

Remark 2. *Several methods can be used to estimate the regression coefficients. For instance, we can simply use central values for b and slopes for a as follows*

$$\tilde{b} = \text{logit}(\widehat{p}(0)), \qquad \tilde{a} = \text{logit}(\widehat{p}(1)) - \text{logit}(\widehat{p}(-1)).$$

We can also use the relation

$$\text{logit}\big(p(y)\big) \quad = \quad b + ay,$$

to estimate the coefficients by ordinary least squares, say $\widehat{a}$ and $\widehat{b}$. Table 1 contains these estimates obtained from first and second approximations based on 1000 samples of size $n = 100$.

	(a, b)	$\widehat{b}$	$\tilde{b}$	$\widehat{a}$	$\tilde{a}$
First order	$(-1, 1)$	0.43 (0.32)	0.82 (0.21)	-0.76 (0.35)	-1.76 (0.87)
Second order	$(-1, 1)$	0.25 (0.61)	0.96 (0.39)	-0.81 (0.34)	-1.51 (0.76)
First order	$(1, 1)$	0.43 (0.32)	0.81 (0.20)	0.77 (0.38)	1.76 (0.87)
Second order	$(1, 1)$	0.12 (0.63)	0.94 (0.36)	0.78 (0.33)	1.54 (0.78)

Table 1: Estimates and their standard errors (in brackets) based on 1000 samples of size $n = 100$.

4.2 Test procedure

We now proceed with the study of the random character of the missingness mechanism. We consider the model given by (1) where Y is $\mathcal{N}(0, 1)$ distributed. We want to test the independence between Y and W. The hypotheses can be reformulated as

$$H_0 : p = w \ (= \mathbb{E}(W)), \qquad H_1 : p \neq w.$$

or equivalently,

$$H_0 : a = 0, \qquad H_1 : a \neq 0.$$

We chose $b = -1, 0, 1$ and we considered alternatives with $a = -10, -1, 1, 10$. We assume that the distribution of Y is known to compare the proposed data driven statistic to the Kolmogorov-Smirnov (KS) one. KS statistic is based on the observed Y (when $W = 1$) which follow the normal distribution under H_0. According to Assumption (A1), we fixed $k(n) = 2$ for $n = 30, 50$ and $k(n) = 3$ for $n = 100$. We chose a theoretical level $\alpha = 5\%$. Empirical levels were very close to the asymptotic 5% and we omitted their values.

Figures 4-7 show the empirical powers equal to the number of rejects of H_0 divided by 1000 (the number of replications).

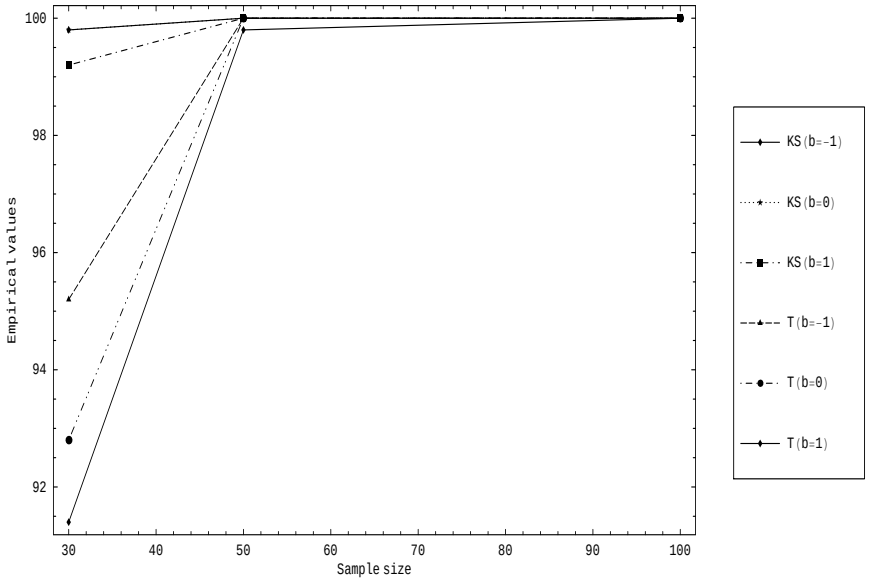

Figure 4: Empirical powers when H_0 coincides with logistic coefficient $a = 0$ and $b = -1, 0, 1$, and when alternatives coincide with $a = -10$. In the legend KS denotes the Kolmogorov-Smirnov statistic and T denotes our data driven statistic. Sample sizes are $n = 30, 50, 100$. The theoretical level is $\alpha = 5\%$.

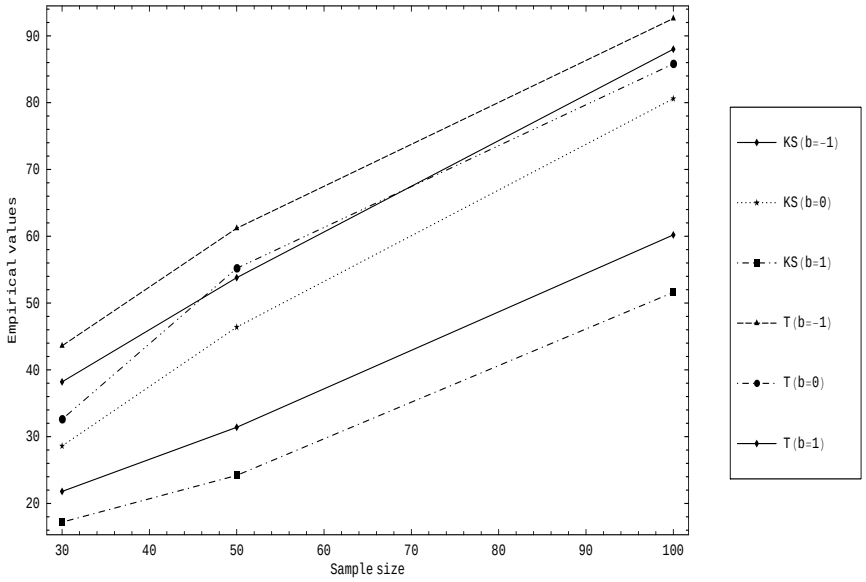

Figure 5: Empirical powers when H_0 coincides with logistic coefficient $a = 0$ and $b = -1, 0, 1$, and when alternatives coincide with $a = -1$. In the legend KS denotes the Kolmogorov-Smirnov statistic and T denotes our data driven statistic. Sample sizes are $n = 30, 50, 100$. The theoretical level is $\alpha = 5\%$.

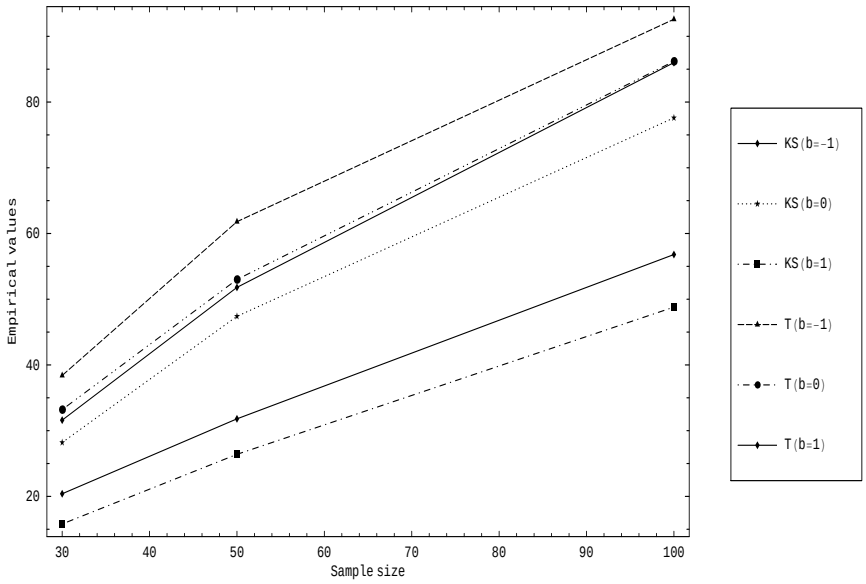

Figure 6: Empirical powers when H_0 coincides with logistic coefficient $a = 0$ and $b = -1, 0, 1$, and when alternatives coincide with $a = 1$. In the legend KS denotes the Kolmogorov-Smirnov statistic and T denotes our data driven statistic. Sample sizes are $n = 30, 50, 100$. The theoretical level is $\alpha = 5\%$.

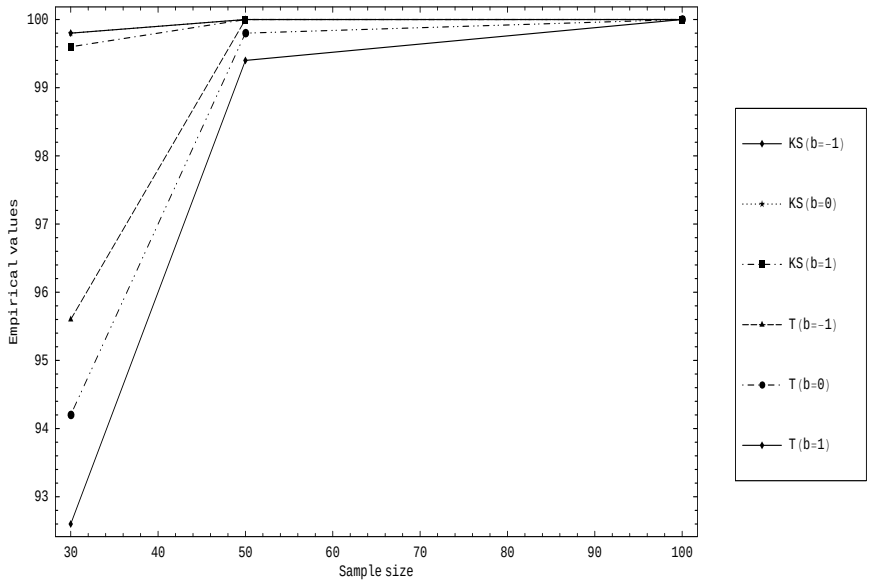

Figure 7: Empirical powers when H_0 coincides with logistic coefficient $a = 0$ and $b = -1, 0, 1$, and when alternatives coincide with $a = 10$. In the legend KS denotes the Kolmogorov-Smirnov statistic and T denotes our data driven statistic. Sample sizes are $n = 30, 50, 100$. The theoretical level is $\alpha = 5\%$.

It can be observed that the Kolmogorov-Smirnov statistic gave slightly better results for large values of $|a|$ and for small sample size. In these cases the missing probability increase or decrease very quickly. Conversely, the data driven statistic provided better results for $|a| = 1$, when missing values are less asymmetric, that is not concentrated on a region of the support $\mathcal{S}$. The case $a = 0.1$ gave empirical powers less than 10 percents for both statistics and we omitted this graph.

Remark 3. *As pointed out in § 3.2 we do not need the full knowledge of the distribution Y but only the $k(n)$ first moments. Under these assumptions we obtained similar empirical powers than those presented in Figures 4-7 where the distribution of Y is known. This illustrates the advantage of the method and the good behavior of the test when the information on Y is partial.*

5 A REAL DATA

The French National Institute of Statistics and Economic Studies (INSEE) provides general information on the population, and in particular on the total incomes (per year) of people in France, including parameters such as mean, standard deviation, median. The IRIS data set contains such indicators on the incomes in 2003. In particular, the mean estimated from this survey was $m = 12713.5$ euros and the standard deviation was $s = 13931$. We are then in position to apply the test statistic with two known moments of the income Y.

We consider the following data: From February to April 2003, the INSEE conducted a survey on the *identities* of the population. Roughly speaking, identity refers to the way to build a place in the French society. The income Y is one of the variables in this survey. We are interested in the probability of missing values concerning Y. There were 380 missing values over 8403 observations. The value 0 here corresponds to people answering they do not known their income. The test can be applied to decide if the missing depends of the income. Despite the large sample size $n = 8403$ the data driven statistic is used with $k(n) = 2$ since it cannot exceed the number of known moments. Applying our test procedure we obtained a p-value less than 10^{-16}. Returning to the data it seems that low incomes are more often missing leading to a greater mean and, by asymmetry of the income distribution, a greater variance.

References

Abramowitz, M. & Stegun, I.A. (Eds.) (1972). *Orthogonal Polynomials.* In *Handbook of Mathematical Functions*, 9th printing. Chap. 22. New York: Dover.

Härdle, W. (1992). *Applied Nonparametric Regression.* Cambridge Books, Cambridge University Press.

Ignaccolo, R. (2004). Goodness-of-fit tests for dependent data. *Journal of Nonparametric Statistics*, 16, 19–38.

Kallenberg, W.C.M. & Ledwina, T. (1995). *Annals of Statistics*, 23, 1594–1608.

Ledwina, T. (1994). Data-driven version of neymans smooth test of Fit. *Journal of the American Statistical Association*, 89, 1000–1005.

Neyman, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20, 149-199.

Rayner, J.C.W. & Best, D.J. (1989). *Smooth Tests of Goodness of Fit.* Oxford University Press, New York.

Rayner, J.C.W. & Best, D.J. (2001). *A Contingency Table Approach to Nonparametric Testing.* Chapman and Hall/CRC.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Rubin, D.B. (1987). *Multiple Imputation for Non-response in Survey.* J. Wiley and Sons, New York. Wiley.